Project Report : Machine Learning(CAP 6610) Object Detection

Vikas Chaubey
UFID : 3511 5826
Email : vikas.chaubey@ufl.edu
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, Fl 32608

Abstract

The aim of this project is to implement a machine learning model which can perform Object detection within given digital image.This machine learning model is implemented using state of the art machine learning approaches such as deep learning neural networks.These neural networks can perform Object detection which is comprised of two different tasks which are image classification and locating the identified object within the image by drawing a rectangle around the object.

1 Introduction

Object detection is a generic term.Actually object detection could be defined as combinations of computer vision tasks which allow a machine or a computer to identify presence of objects of specific class in visual imagery.Given multiple classes of different objects a classification algorithm can identify the type of class for a given image.But object detection algorithms combine this task of class identification with the identification of spatial location of the identified object in the given image.hence any object detection algorithm perform two separate tasks.

A general example of object detection :

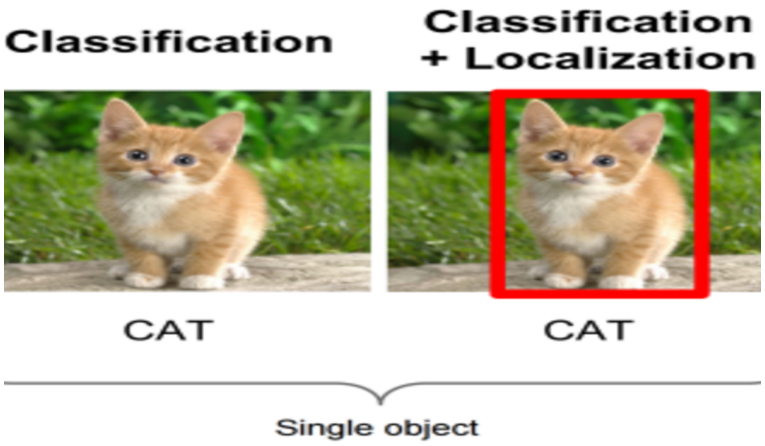

Figure 1: Object detection Within a image.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

1.1 Motivation

The main reason and motivation behind choosing object detection for my final machine learning project is that computer vision as a field has always fascinated me. The wide range of use cases in different fields makes this domain really interesting for example object detection is currently being used to build self-driving cars (built by Tesla), Delivery drones (being developed by Amazon), face recognition and Pedestrian tracking etc.The concepts that i had learned in the class such as feature engineering, classification problems, model optimization, regularization, model over fitting , activation functions have helped me immensely to implement this project.Also i got chance to explore the field of object detection within images in general that includes exploring traditional object detection approaches to new machine learning based methods and finally deep learning based effective object detection solutions. Working on this project has provided me with good experience with cutting edge technologies being used in machine learning domain in research and industry as well.

2 Formulation of statement :

In this project solution for object detection problem is implemented using deep learning neural networks.In order to do image classification convolution neural network is used.This convolution neural network is trained to identify five different classes or categories (five different types of animals).The second part of the problem which is locating object spatial location within the image is resolved using class activation mapping (Grad-Cam) algorithm.Using this approach the model does not require data which is labelled with bound boxes for training in order to learn the identification of spatial location of the objects.Class activation mapping is a technique which utilizes the output from last convolution layer of convolution neural network to identify the spatial location of object within a digital image.hence the model needs to be trained only using labelled data(class names only).The problem formulation is focused on classification.

The convolutional neural network performs convolution operation.The convolution layer has multiple filters(f) also called kernels,kernels are convolved with image Input, the convolution operation could be defined as :

$$C[m,n] = (I * f)[m,n] = \sum_i \sum_j I[i,j] f[m-i, n-j] \tag{1}$$

Forward propagation : in the forward propagation the convolution output from previous layer(A) is convolved with tensors(W) and an intermediate value (Z) is calculated for current layer by adding bias(b) to the resultant.Then this intermediate value is fed to the activation function(g), and this outcome is forwarded to the next layer.here [l] represents the current layer and [l-1] represnets previous layer.

$$Z^l = W^l * A^{l-1} + b^l \tag{2}$$

$$A^l = g^l(Z^l) \tag{3}$$

Back propagation in Convolution Layers: after complete pass of forward propagation, flow is executed backwords to calculate derivatives in order to obtain weights.

$$dA^l = \frac{\partial L}{\partial A^l} \quad dZ^l = \frac{\partial L}{\partial Z^l} \quad dW^l = \frac{\partial L}{\partial W^l} \quad db^l = \frac{\partial L}{\partial b^l} \tag{4}$$

final step involves calculating dZ which could be calculated as :

$$dZ^l = dA^l * g'(Z^l) \tag{5}$$

Here L represents the loss function(categorical loss or log loss) which has to be minimized , for multi label classification.

$$L(y, \hat{y}) = \sum_{j=0}^{M} \sum_{i=0}^{N} (y_{ij} log(\hat{y}_{ij})) \tag{6}$$

Visual explanation of forward and backward propagation:

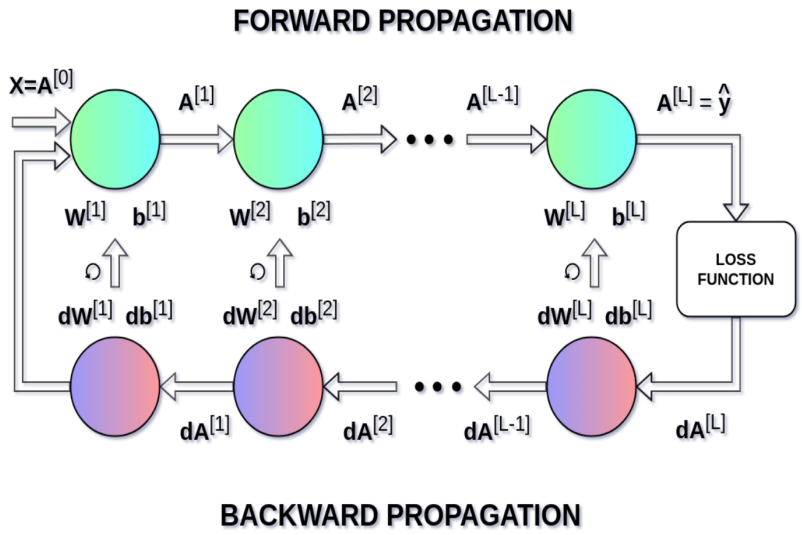

Figure 2: Visual Representation of forward and Backward Pass

3 Algorithms Used to Solve Object Detection Problem:

3.1 Convolution Neural Network(CNN) :

Convolutional neural networks are the most popular neural network for processing image data.Convolutional neural networks perform better than other traditional image processing techniques as well as they also outperforms traditional machine learning methods in image processing for object detection.The main reason behind their good performance is that the architecture of the neural network is inspired by the functioning of part of human brain which is responsible to process visuals seen by eyes.When a visual is seen by eyes it triggers a specific portion of brain , different type of stimulus triggers different layers of that brain portion.This is the exact concept which is used in convolution neural networks.Also these networks perform better than feed forward networks which are fully connected because in these networks all the neurons are interconnected to each other as a result when it comes to processing image data especially big images , number of model weights that has to be calculated increase drastically hence such networks perform poorly with images on the other hand in case of convolution neural network all every neuron is interconnected hence that reduces the load of weight calculation still being accurate for image processing because it follows the exact processes the human brain follows in order to process visual data.

Convolutional neural networks ingest and process images as tensors, and tensors are matrices of numbers with additional dimensions.For example in case of an image these networks has to be fed the matrix representation of the image(width, height, depth), here height and width of the image matrix represents the rows and columns of the matrix while the depth represents three channels present in the Image which are Red, Green , Blue.

Definition of Convolution: “to convolve” means to roll together.A convolution process between two functions is the integral measuring how much two functions overlap as one passes over the other.Convolution layers make use of this concept instead of traditional matrix multiplication.

Architecture of convolution neural network mainly comprised of follwoing layers:

1) Convolution layers.
2) Pooling layers.
3) Dense Layers(Fully connected layer).

1) Convolution Layer: The convolution layer is comprised of many filters also called as kernels of certain dimensions, in this layer these filters are convolved with image input and class activation maps

are generated.number of class activation maps depend upon the number of filters present in this layer. with the help of those class activation maps the convolution layers learn the features of visual data.

2) Pooling Layers : The class activation maps generated in the convolution layer are fed to the pooling layers.this layers perform the function of max pooling, down sampling and sub sampling.For example in max pooling sampling is done in such a way that in these layers Only the locations on the image that showed the strongest correlation to each feature are preserved, and other values are ignored.Subsequently the outcomes of this layer are fed to the next convolution layer.

3) Dense layers : final outcomes or stacked class activation maps are finally fed to the feed forward network which is fully connected and this network performs the classification by using the data and an activation function.

A simple Block Diagram of convolution neural network containing multiple layers :

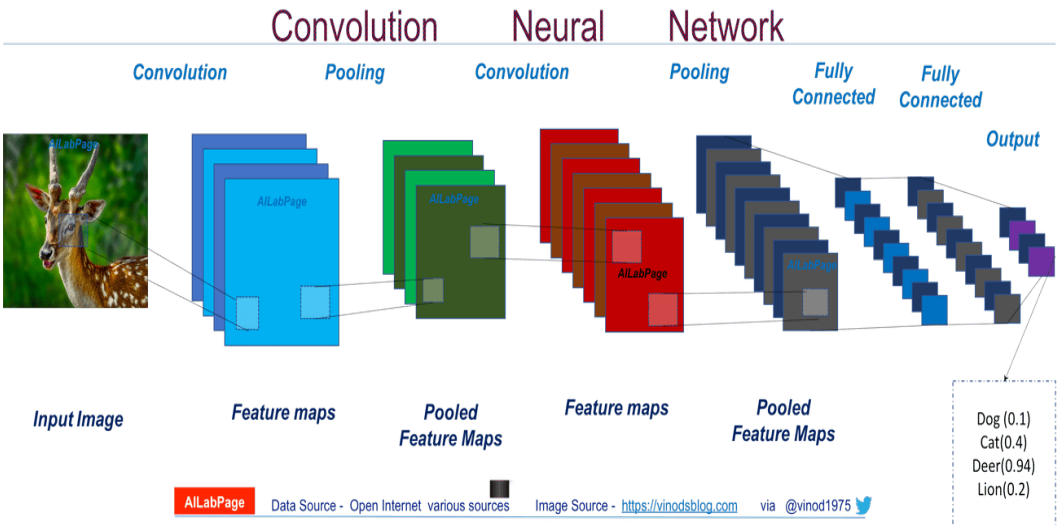

Figure 3: Visual Representation of Convolution neural network

3.2 Gradient-Weighted Class activation mapping (Grad-Cam Algorithm) to identify object location

Once the CNN performs the classification of the input image the next problem is to locate the classified object within the image.This is implemented using the Grad-Cam algorithm which based on class activation maps.The Grad-Cam algorithm makes use of the outcomes of the last convolution layers of the CNN for the given input image, it uses the class activation maps of the classified category and input image and generate the heat maps for the identified objects .These heatmaps could be used to generate the bounding box around the spatial location of the object.This does not require to train the model explicitly to locate identified objects within given image.

Visual Explanation of Grad- Cam functionality :

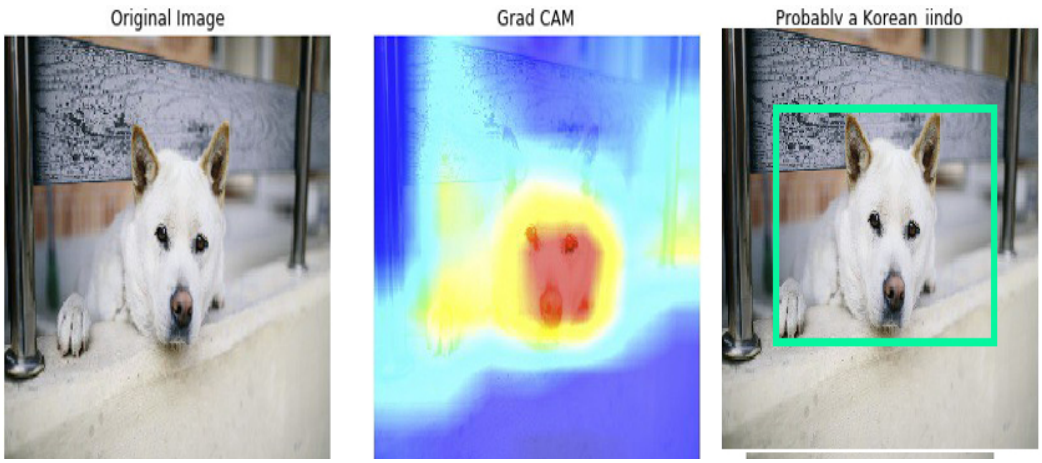

Figure 4: input image, heatmap generation, Locating the Dog within boudning Box

Another example with multiple object detection :

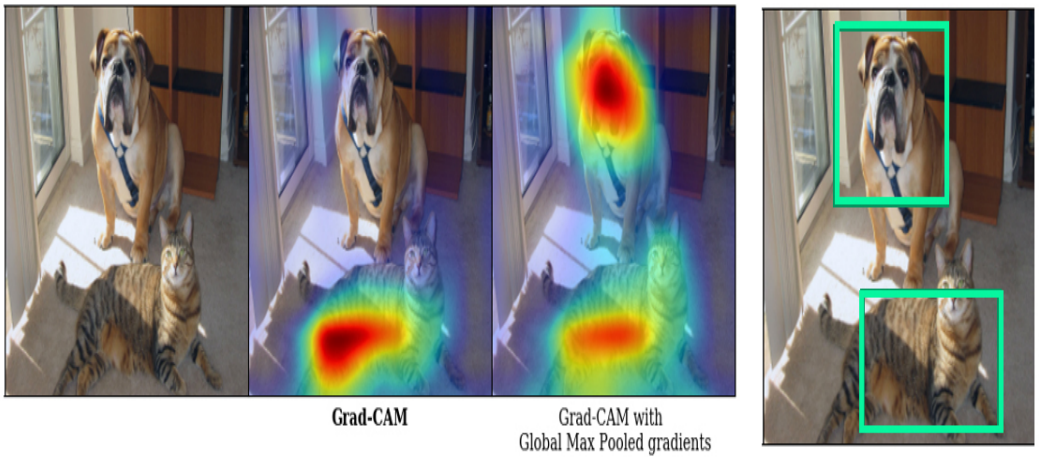

Figure 5: input image, heatmap generation, Locating the cat and dog within boudning Box

4 Project implementation :

The project is implemented using python 3.7 and keras library.To implement Grad-Cam algorithm Open CV library is used for image processing.

4.1 Data Set Preparation

4.1.1 Proposed dataset (Cifar-10):

As per project plan proposed during project proposal the idea was to train the classifier using Cifar-10 data set for training and testing of the model. This dataset is for image classification. It consists of 60,000 images of 10 classes (each class is represented as a row in the above image). In total, there are 50,000 training images and 10,000 test images. The dataset is divided into 6 parts – 5 training batches and 1 test batch. Each batch has 10,000 images.

4.1.2 The problems with Cifar-10 dataset :

After training the classifier model on cifar-10, the outcomes on test data were highly inaccurate the problem was the very small size of images in data set, cifar-10 data set is comprised of 32 X 32 pixel images, on a display screen this size is equivalent to a desktop thumbnail, also the images are very low quality which makes the images visually blurred, hence it was very difficult to implement localization and bounding boxes on such low quality small images. Also model did not generalize well when tested with bigger test images randomly taken from outside the data set.Hence there was a need to change the data set for training.

4.1.3 Data set preparation using imagenet :

The project was implemented using data set prepared by Imagenet images.This new data set consists of five different classes of animals which are Dog,Cat,Mouse,bird and Frog. Total 7500 images were downloaded across all 5 classes , 1500 images for each class(Dog,Cat,Mouse,bird and Frog)and this data set was finally used to train the project model.

4.2 Training ,Validation and Accuracy

among all the data set images 80 percent of the images are used for training remaining 20 percent images are used for validation testing of the model.So total 6500 images were used for training and remaining 1000 images were used for validation for the model.This ios the number of images across all five classes devided evenly in both training and validation set. The model is trained on MacOs platform using local GPU (Radeon Pro 560X 4 GB).the model is trained for total 50 epochs.The total training time for the model was approximately 43 hours.

1) The accuracy obtained for model on the training data set is nearly : 77 percent and The accuracy obtained for model on the validation data set is nearly : 55 percent.
The Accuracy Plot for Train/validation datasets is as below :

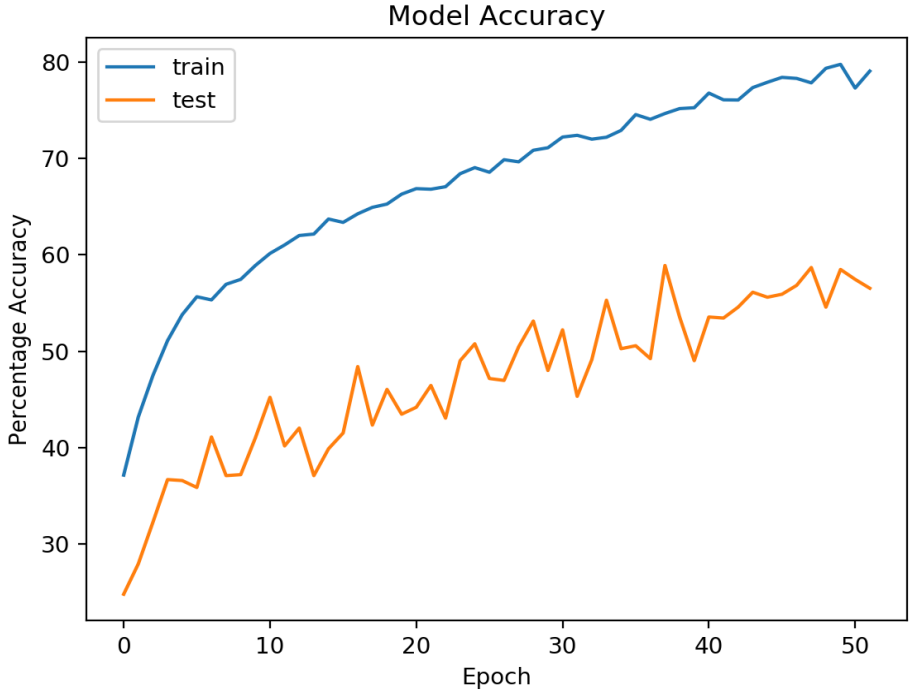

Figure 6: Model Accuracy

2) The loss obtained for the model on the training data set is nearly : 0.56 and The loss obtained for the model on the validation data set is nearly : 1.56

The Plot for Train/test datasets is as below :

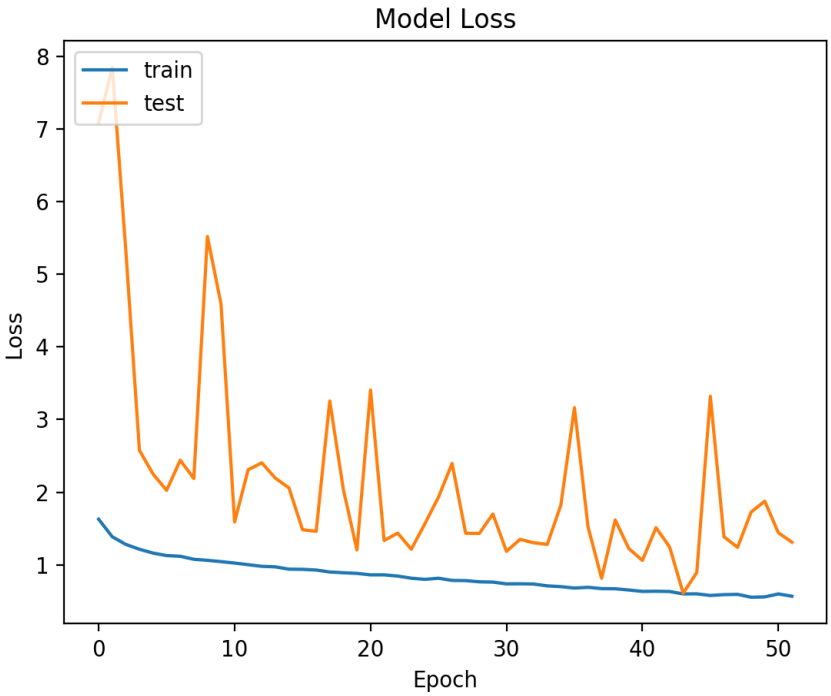

Figure 7: Model Loss

4.3 Testing Project functionality on some images

Some of the testing results performed on the classifier are as follow:

1) Accurately predicted and bound boxed image of a Dog :

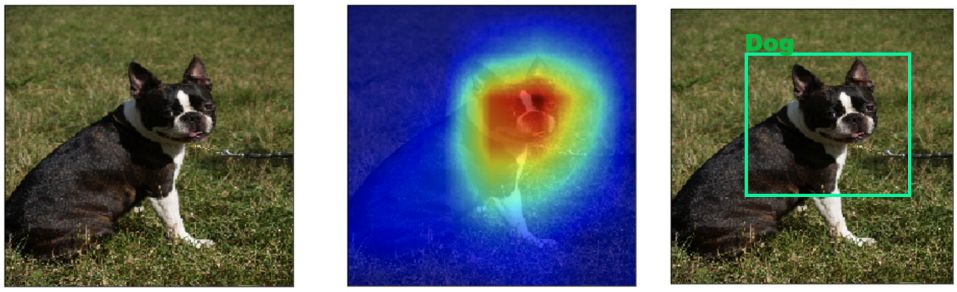

Figure 8: Dog Image, Heat Map generated by Grad Cam, Classifier output and Boudning Box created using heat Map

2) Accurately predicted class But inaccaurate prediction of location :

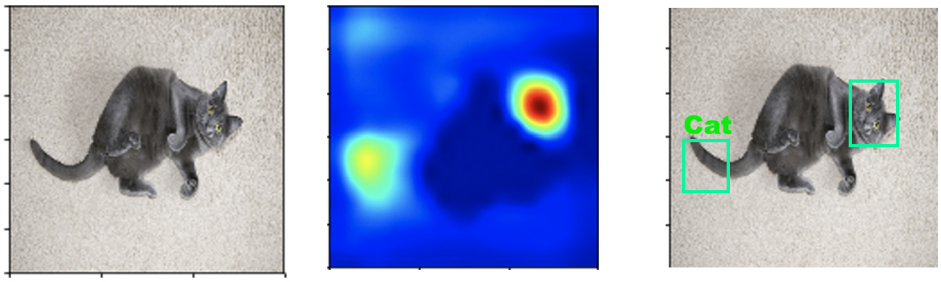

Figure 9: cat Image, Heat Map generated by Grad Cam, Classifier output and Boudning Box created using heat Map

3) Accurately predicted class But inaccaurate prediction of location :

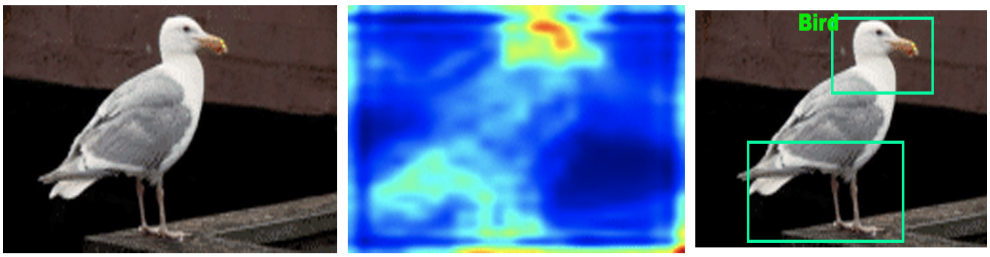

Figure 10: bird Image, Heat Map generated by Grad Cam, Classifier output and Boudning Box created using heat Map

5 Conclusions :

After implementing object detection using convolution neural network (For classification) and Grad-Cam(For identification of object location) and testing the application model with multiple images the following conclusions were made:

1)classifier model is generalizing well on image data out of training and validation sets hence the classifier produces good accuracy while doing prediction of classes

2) After classification, the object localization performed by Grad Cam algorithm using class activation maps from the last layers of convolution neural networks is not very accurate, on testing with multiple images and different classes results were found to be very inaccurate.

3) It is true that using this approach the effort to prepare large data sets for object localization training of neural network could be reduced significantly but in order to measure the accuracy of bounding boxes annotated data is required otherwise accuracy of bounding boxes can not be determined, it could be determined by human intelligence but again that will take significant amount of effort.

4)in order to produce high accuracy for both classification and localization, The convolution neural network has to be trained thoroughly with large data sets, A very accurate classifier will produce better class activation maps of image features which will be used by Grad Cam effectively to produce good localization result as well.Hence localization done using Grad -cam highly depends on the Classifier accuracy.

5) Other object detection algorithm like YoloV3 are still much faster and more accurate than object detection implemented using Class activation maps.

References

[1] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, pages 3431–3440, 2015.

[2] A. Bergamo, L. Bazzani, D. Anguelov, and L. Torresani. Self-taught object localization with deep networks. arXiv preprint arXiv:1409.3964, 2014. 1, 2

[3] P . Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

[4] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian,"Grad-CAM++: Generalized Gradient- Based Visual Explanations for Deep Convolutional Networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV , 2018, pp. 839-847.

[5] Francois Chollet. 2017. Deep Learning with Python (1st. ed.). Manning Publications Co., USA.

[6] R. L. Galvez, A. A. Bandala, E. P. Dadios, R. R. P. Vicerra and J. M. Z. Maningo, "Object Detection Using Convolutional Neural Networks," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 2023-2027.

[7] J. Huang, A. Fathi, V. Rathod, I. Fischer, C. Sun, Z. W ojna, M. Zhu, Y . Song, A. Korattikara, S. Guadarrama, K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors", pp. 1-21, April 2017

[8] keras. https://keras.io/

[9] Open CV. https://opencv.org/

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[11] [33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In ECCV, 2014.

[12] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In ICLR Workshop, 2015.